Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies

Daniel Whalen

Anthony Pepitone

Linda Graver

Jon D. Busch

July 2001

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Substance Abuse and Mental Health Services Administration
5600 Fishers Lane
Rockville, MD 20857

Acknowledgments

This report is compiled from a larger study, which itself is the result of substantial contributions by numerous people and organizations. Linda Graver of the MEDSTAT Group coordinated the project and monitored the results. Daniel Whalen, Anthony Pepitone, Jon Busch, Linda Graver, and Nancy Jordan drafted and reviewed the report. Anthony Pepitone, Jon Busch, and Domingo Arenas provided technical guidance. Joan Dilonardo, Jeffrey Buck, and Mady Chalk guided the work and provided many helpful comments and suggestions throughout this project.

Disclaimer

The MEDSTAT Group prepared this report for the Substance Abuse and Mental Health Services Administration (SAMHSA) of the U.S. Department of Health and Human Services under Contract No. 270-96-0007 to Joan Dilonardo, Ph.D., Government Project Officer. The content of this publication does not necessarily reflect the views or policies of SAMHSA, nor does it necessarily reflect the views of any of the Advisory Panel members. The authors are solely responsible for the content of this publication.

Public Domain Notice

Recommended Citation

Whalen D, Pepitone A, Graver L, Busch J.D. Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies. SAMHSA Publication No. SMA-01-3500. Rockville, MD: Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, July 2000.

Electronic Access and Copies of Publication

This publication can be accessed electronically through the following Internet World Wide Web connections: http://www.samhsa.gov and http://www.ncadi.org. For additional paper copies of this document or associated background reports, please call the National Clearinghouse for Alcohol and Drug Information, 1-800-729-6686.

Originating Office

Office of Managed Care, Center for Substance Abuse Treatment, 5600 Fishers Lane, Rockwall II Building, Suite 740, Rockville, MD 20857

Executive Summary

This report describes the concepts behind record linking and the specific application of record linking in building databases integrating information about mental health (MH) and alcohol/drug (AOD) services. The MEDSTAT Group constructed these databases as part of a contract (270-96-0007) sponsored by SAMHSA's Center for Substance Abuse Treatment (CSAT) and the Center for Mental Health Services (CMHS). Each Integrated Database (IDB) includes comprehensive information for MH and AOD services from both the MH and AOD State Agencies, as well as Medicaid Agencies for three States: Delaware, Oklahoma, and Washington.

A variety of methods can be employed to link records from different data sources and these methods vary in terms of complexity, efficiency, and accuracy. Simple matching and deterministic methods are useful for certain applications, and while these methods are relatively simple to implement, they can also produce inaccurate results. By contrast, probabilistic linking methods are relatively complex, but tend to produce more accurate results. The theoretical underpinnings of various approaches to record linkage are discussed, along with the relative strengths of each approach.

Probabilistic linking routines were developed for use in combining Medicaid data with MH/AOD agency data for three States. The nature and function of these routines are described in light of the experience gained in processing State data. Results suggest that when compared with other record linkage methods, probabilistic matching produces more links than other methods and that many of these links are missed by other methods. This indicates probabilistic linking routines are more accurate than other routines for matching person-level data.

To facilitate dissemination of these linking routines, the source code used in the linking process is disseminated at no cost via the project web site. Potential applications and extensions of this methodology are discussed and future directions are outlined.

Table of Contents

Tables

Figures

Chapter One: Overview

The primary purpose of this document is to describe the innovative linking methodology that is being developed and applied on the Integrated Data Base (IDB) project. We begin by describing the context of the linking methodology: the project objectives, the characteristics of source data, and the structure and content of the databases we are building. Next, we review standard methods of data linking and identify their strengths and weaknesses with respect to the IDB project. Subsequent chapters detail the IDB probabilistic linking process, describe how it is applied to our unique data sources and evaluate the degree to which it has been successful. The paper concludes with a discussion of potential improvements and includes references and a description of the source code for all steps of the linking process.

CSAT, CMHS and The MEDSTAT Group (prime contractor on the project) are working collaboratively with the three States to develop databases that integrate MH, AOD and Medicaid program data. The goal of the project is to create State-specific databases in which we identify and organize, at the person level, data on all services that a person receives, across all providers. IDB databases provide CSAT, CMHS, and the collaborating States with data for evaluating MH/AOD use and relative cost at the person level. The IDB includes information for all persons who have received MH/AOD services (or services potentially MH/AOD related) either from the MH/AOD agency or from Medicaid providers. In addition, the IDB contains information about all Medicaid services for this MH/AOD population, including medical services not related to any MH/AOD condition.

The IDB project is a secondary data development effort. Secondary data, such as administrative healthcare data, are invaluable research tools. As illustrated by Romano and Luft (1992), secondary data provide comprehensive data at relatively low cost to researchers, and "have important roles in studies of appropriateness, quality, and outcomes of care." A major strength of administrative data is inclusiveness. These data are collected unobtrusively and systematically from entire groups as part of administrative goals, such as claim adjudication or utilization monitoring. By conducting a secondary data development project, we are able to leverage the significant investments of time and effort that have occurred collecting and organizing the primary databases that constitute our source data; the time and resources necessary to collect comparable primary data would be prohibitive. At the person level, costs for secondary data are a small fraction of the cost for gathering equivalent data.

Sources of secondary data include private entities such as clinics and hospitals, State agencies, and Federal bureaus and programs such as Census, and Medicare data. Data sources for the IDB consist of secondary data from State agencies collected for various purposes. The MH/AOD Agency data are collected primarily to track service use and patient characteristics, whereas the principal purpose of Medicaid data is the adjudication of claims. As might be expected, the structure of data from these sources reflects the variety of purposes for these data. MH/AOD Agency data generally provide more patient information than do Medicaid data. Medicaid data, however, include more service information, including charge and payment data.

One of the challenges in using secondary data for research is that the data are not collected for purposes relevant to the particular research question(s). For example, Medicaid data are collected primarily to adjudicate claims for treatment provided for eligible persons. Thus, Medicaid data might be able to answer questions relevant to patterns of utilization and costs of particular types of services for persons eligible for Medicaid. With the exception of age, gender, and race/ethnicity, Medicaid data contain little information about patient characteristics. Thus, if the researcher is interested in how clinical characteristics may influence the receipt of particular types of services, it is unlikely that Medicaid data provide appropriate information.

A similar challenge is presented by secondary data sets that may contain information relevant to a particular research question, but the information produces an incomplete answer. One example of this relates to public costs for mental health and alcohol and other drug treatment. While it may be useful to estimate the per person costs of providing mental health services through a State agency, it is not uncommon for mental health and alcohol and drug State service records to be maintained in different data systems within a State. Even within a relatively integrated State agency or department, information about client characteristics, utilization and costs may be organized and stored on multiple files. Although it is likely persons within a system will have a unique identifier within a given system, that same identifier may not be in other systems. For example, a Department of Motor Vehicles might use drivers' license numbers to uniquely identify people, but this number is not likely to be recognized by the same State's Medicaid agency. Moreover, the Medicaid ID is probably meaningless at the DMV.

Analyses of per person costs for mental health and alcohol and drug treatment services conducted on either file segment would yield only a partial answer about the true State costs. Further, since both mental health and alcohol and drug services may also be delivered as Medicaid services, State agency costs of delivering mental health and alcohol and drug treatment services would represent only a partial answer to the question of public costs for mental health and alcohol and drug treatment.

It may be possible to overcome some of these limitations by linking together files from disparate sources. For example, linkage of State mental health and alcohol and drug treatment services records with Medicaid records allow for increased precision in the estimation of utilization and costs on a per person level. Similarly, linkage of various record types, each with their own characteristics, may provide an enriched data file for persons who are present in each file. One of the greatest challenges involved in the creation of linked files is the absence of one unique identifier for each person across each of the electronic data files.

One of the goals of the IDB was to develop procedures for accurately linking person-level records from multiple files and to do this with a minimum of human intervention. Probabilistic record linking techniques were used to combine MH/AOD Agency data and Medicaid data for the IDB. These techniques use multiple, sometimes conflicting, criteria to resolve record links accurately.

Automation of record linkage was pioneered by Howard Newcombe (Newcombe, 1967; Newcombe, Kennedy, Axford, & James, 1959) and later advanced by two Canadian statisticians, Ivan Fellegi and Alan Sunter, who formally described the mathematical model of probabilistic techniques (Fellegi & Sunter, 1969). This model remains the principal technique used in record linking. Fellegi-Sunter methods are used today by the U.S. Bureau of the Census, the Department of Agriculture, Statistics Canada, and the National Cancer Institute (Weber, 1995). This report examines the process developed and the results of that process.

The IDB provides a research-ready, empirical database to CSAT, CMHS, and three participating States. Each database makes possible the analysis of costs and utilization for programs and conditions in the aggregate and at the person level. It also allows researchers to track health care usage for individuals between State MH/AOD agencies and Medicaid providers. To create this functionality required a mechanism to relate data from multiple and disparate sources without requiring a common identifier. This mechanism is at the heart of the construction of the integrated database and IDB-based analyses.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter Two: The Integrated Database

The Integrated Database (IDB) project is sponsored by SAMHSA's Center for Substance Abuse Treatment (CSAT) and the Center for Mental Health Services (CMHS). An ongoing mandate for CSAT and CMHS is to direct efforts to evaluate the organization and financing of care for persons with MH/AOD conditions. SAMHSA recognizes that efforts to improve the delivery and financing of MH/AOD services have been deterred by a lack of comprehensive information on the cost and utilization of MH/AOD services, the impact of MH/AOD conditions on individuals' overall health status, and the outcomes of providing MH/AOD services.

Project Description and Objectives

Broadly stated, the principal goal of the IDB project is to develop databases for each of three participating States (Washington, Oklahoma and Delaware) that integrate data from State Medicaid programs with data from State and local mental health and substance abuse agencies. State agencies typically have the ability to look at costs and service utilization only within their own agency. This project breaks new ground by performing large-scale linking of individuals' services across agencies and across service providers. The IDB provides SAMHSA and collaborating States with a research-ready resource for evaluating the utilization and cost of comprehensive health services at the person level for individuals with MH/AOD conditions. By incorporating Medicaid and MH/AOD services, the IDB enables investigation of the dually-diagnosed as well as an exploration of the impact of MH/AOD conditions on other medical comorbidities and vice versa. Ultimately, we expect this data development effort will provide CSAT/CMHS with an empirical base for MH/AOD policy development.

A second goal of the project is to create a model, or framework, that can be used by other States for their own data integration efforts. To this end, MEDSTAT has developed algorithms and computer programs that perform person-level record linking from multiple data sources for the IDB. By making these methods available, other States may embark on similar projects and benefit from the experiences of the original participants. Lessons learned in the implementation of the proposed methodology will be transferable to other States. Creation of the database involved the application of sophisticated linking routines to the data. These routines allowed us to link records for persons whom we might have otherwise concluded represent different individuals. Early feedback from the States participating in the project is that information derived from the IDB is extremely useful and interesting to State policy makers.

Data Sources

The IDB project combines substance abuse and mental health data from each State's MH/AOD agencies with State Medicaid program data. These two data sources provide very different data, often for the same individuals. MH/AOD Agency data describe recipients of services and their conditions. In general, MH/AOD Agency data contain rich descriptive client information, but relatively little information concerning services. The primary purpose of Medicaid data is to facilitate the adjudication and payment of claims. Accordingly, Medicaid data include detailed information concerning services, but relatively little information about the client.

Medicaid Data

Two types of Medicaid data files are used to construct the IDB: files of client eligibility data and files of claims and encounters. All Medicaid files use a patient ID key (Medicaid ID) for relating data, and Medicaid data files from each State are structured similarly. Eligibility data provide client information for persons eligible for Medicaid services. Identifying data, demographic information, and time-based eligibility information for Medicaid clients are extracted from these eligibility files. Each file typically contains one record for each eligibility period per person, and is designed to be used for determinations of eligibility status. The detail of Medicaid covered services is recorded in fee-for-service (FFS) claims and managed care encounter files. These are the largest files processed by the project team. The FFS and encounter files include records for outpatient and physician services, hospital, skilled nursing, other institutional claims and prescription drugs.

The following agencies provided Medicaid data for the project:

- Delaware: The Department of Health and Social Services, Division of Medicaid,
- Oklahoma: The Oklahoma Health Care Authority,
- Washington: The Department of Health and Social Services, Medical Assistance Administration.

AOD and MH State Agency Data

Each of the three States administers MH and AOD services differently. In Oklahoma, a single agency is responsible for MH/AOD services, while in Delaware and Washington there are two agencies. The two Delaware agencies administer both MH and AOD; one agency is responsible for providing services to children, whereas the other agency is responsible for providing services to adults. By contrast, Washington administers MH services through its mental health agency with AOD services being administered through a separate agency. Such administrative arrangements have specific implications for how data are segmented and stored.

Delaware

Delaware MH/AOD services are provided by two separate agencies, one dealing with children and youth, and the other agency has responsibility for adults. Data from these agencies are distributed across more than 20 different data files. The Department of Services for Children, Youth and Their Families, Division of Child Mental Health, supplies MH/AOD Agency data for children and adolescents in a combined demographics and services file. Adult MH/AOD Agency data are supplied by the Delaware Department of Health and Social Services, Division of Alcoholism, Drug Abuse, and Mental Health. The Division of Medicaid, in this same Department, supplies the Medicaid data for Delaware. Adult MH/AOD Agency data are supplied in a variety of different formats:

- Community MH data are provided as separate client and services files,
- State MH hospital data are supplied as separate admission/discharge and monthly activity data,
- AOD data are supplied as a combined client and episode file.

Oklahoma

One agency in Oklahoma is responsible for both AOD and MH services. The Department of Mental Health and Substance Abuse Services provides three separate files: a client demographic file, an admission file, and a services file. The services file contains information on AOD treatments, community MH services, institutional MH services, and outreach programs.

Washington

Two divisions are responsible for MH/AOD services in Washington, one for AOD services and another for MH services. Both divisions are part of the Department of Social and Health Services, as is the State Medicaid authority. Washington provided non-Medicaid MH/AOD data to the project in 18 separate files. The Department of Social and Health Services, Mental Health Division, provides separate mental health client and service files for State hospital, institutional, and community services. The Department of Social and Health Services, Division of Alcohol and Substance Abuse provides client demographic data in a file separate from service details.

Unique Clients

To provide a sense of the scale of the IDB project, Table 1 below shows the number of unique clients from each data source before the linking and integration process. Medicaid client counts were obtained from the Medicaid eligibility files. Client counts for MH/AOD agencies were obtained from client files where available or from the combined service files when separate client files were not available.

Table 1: Unique Clients by State and Agency

State	Data Source	Unique Clients
Delaware	Adult MH/AOD	11,369
	Youth MH/AOD	2,131
	Medicaid (All)	101,713
Oklahoma	MH/AOD	119,000
	Medicaid (All)	467,176
Washington	AOD	38,539
	MH	99,908
	Medicaid (All)	1,335,387

Data Confidentiality

Strict confidentiality restrictions apply to person-level healthcare data in general and to the data used on the Integrated Database Project in particular. This is true with records of mental health treatment, and especially alcohol and drug treatments. Healthcare data with identifying information or identifiable data are treated as confidential and access is restricted. Identifying data are any data that may directly identify a person. Names and ID numbers are examples of identifying data. Also considered confidential are *identifiable* data. Identifiable data do not directly identify persons, but may facilitate identification. Date of birth combined with ZIP Code and race constitutes identifiable data because they provide a reference that may identify a person in certain circumstances. Confidentiality constraints represent a potential barrier to the application of person-level linking. Identifiers are necessary for linking person records since record linking actually involves person identification and any data that identify an individual is confidential data.

The Department of Health and Human Services' 42 CFR provisions specify strict confidentiality of alcohol and drug treatment records: records with identifying information can be disclosed only in specified, controlled circumstances. One of those circumstances is for research purposes at the discretion of the Program Director. Data were released to the IDB project for the purpose of research. The 42 CFR regulations are subject to State interpretation – some States interpret them much more stringently than others. To comply with HHS 42 CFR and State requirements, MEDSTAT signed data sharing, non-disclosure agreements with each State and agreed to protect the confidentiality of the data throughout the project.

With certain restrictions, the three participating States released confidential data to MEDSTAT for purposes of building the IDB. A variety of policies and procedures are in place to protect the data, which include:

- All project staff sign confidentiality and non-disclosure agreements that tightly control their use and handling of the data.
- Data are stored and processed only on a dedicated secure server with access only through password protected workstations.
- All data received are tracked, including date of receipt, type of data, time period of the data, supplier and contact person.
- A bonded contractor disposes of printed output when it is no longer needed.

The personal identifiers and other identifying information that are crucial for probabilistic linking are used for intermediate processing only and the final research database does not contain any identifying information. Instead, new encrypted personal identifiers are used to structurally connect the components of the database, replacing identifying data. Finally, before data are returned to a State, MEDSTAT requires an inter-agency agreement from participating parties that permits inter-agency sharing of the data.

Chapter Three: Record Linking

In a simplified view, two general types of linking techniques have been developed: linking techniques (deterministic and probabilistic) and match merging techniques. Match merges are generally simpler to perform, but they often create inaccurate and incomplete matches due to data errors and omissions. In contrast, linking activities tend to be more complex than match merges, but they are able to overcome most data errors and omissions and create more accurate and complete matches. These methods of linking persons across data sources are discussed in more detail in this chapter.

Linking Terminology

Linking can be best described after defining a number of underlying concepts. The terms below apply variously to all linking techniques – including match merging, deterministic linking and probabilistic linking.

Record-pair – a combination of records from two files such that one half of each pair is derived from the first file and the remainder is from the second file.

Links – record-pairs that represent the same person or entity (a.k.a. *linked*). In match merging, the "matched" records are *links*.

Non-links – record-pairs that do not represent the same person or entity.

Decision Space – the complete set of record-pairs that are evaluated to determine links.

Joined Records – the *collection* of record-pairs that make up the decision space. Joined records are the sum of all links and non-links.

Cartesian Product – a set of joined records constructed from two files such that each record from the first file is paired with every record from the second file, as depicted in Table 2.

Table 2: Cartesian Product Record-pairs

		File 2		
	Record	**X**	**Y**	**Z**
File 1	**L**	L-X	L-Y	L-Z
	M	M-X	M-Y	M-Z
	N	N-X	N-Y	N-Z

Blocking – a technique to limit the decision space to a manageable size without eliminating potential links.

Identifying Variables – information that can be used to identify a person. This includes direct identifiers such as name and indirect identifiers such as date of birth and race.

***Comparison Variables* –** identifying variables used in comparing the two halves of a record-pair.

***Comparisons* –** the result of equating comparison variables from a record-pair. Record-pairs typically contain a mixture of comparisons in both agreement and disagreement. Comparisons are part of the process of evaluating record-pairs to determine links.

***Dichotomous Comparisons* –** comparisons that evaluate as either true or false – agreement or disagreement.

***Continuous Comparisons* –** comparisons resulting in a numeric score that reflects partial agreement ranging from complete disagreement to complete agreement.

***Weights* –** numeric values that indicate the overall importance of a comparison relative to other comparisons. The discriminating power of each comparison variable – its importance in determining links – is expressed as a weight.

***Scaling* –** adjusting the weight for a comparison variable to reflect the relative frequency of a specific value.

Score & ***Scoring* –** the sum of the products of all the comparisons with the associated weights. The score is used to evaluate record-pairs and determine links and non-links. When weights are applied and summed into scores, the scores for record-pairs that should be linked are generally higher than scores for the record-pairs that should not be linked.

***Decision Groups* –** the division of the decision space into groups based on scores for the purpose of deciding which records should be linked. Record-pairs can be classified as links, non-links, and uncertain pairs.

Uncertain pairs– record-pairs for which a link or non-link determination cannot be made.

Match-Merge Methods

Match merging is a familiar concept in data processing. This technique uses keys (identical variables) on each file to facilitate the match merge: records from two files are combined when the respective keys from each file are the same. An example of a match merge is the use of a Medicaid ID to combine current Medicaid Eligibility information with Medicaid Claims. A match merge can use a simple, single key, as with the Medicaid example, or use a more complex key made up of several variables.

Identifiers serving as keys generally provide good, but not perfect, match-merge results. Problems may occur due to omissions or errors. Data errors and omissions occur for numerous reasons. Information may be omitted because a patient is unwilling or unable to supply it. For example, people are often reluctant to supply their Social Security Number (SSN). Even when patients supply personal information, it may be recorded incorrectly. Digits may be transposed in IDs, names may be difficult to spell or ordered incorrectly. Data errors may occur when written information is illegible. If the key is missing for a given record, matching that record is not possible. Two outcomes are possible for records with incorrect key values: either the incorrect records do not match with any records on the second file or they may match with the wrong record.

Identifiers issued and managed by a specific agency are routinely available on that agency's data and errors are rare. For example, State Medicaid Agencies use their own Medicaid ID to identify clients and maintain the integrity of data. Consequently, Medicaid IDs are present on all Medicaid eligibility, claim, and encounter records. These IDs are valid and usually accurate. Similarly, State MH/AOD Agencies define and use patient IDs – different from Medicaid IDs – to identify their clients and maintain the integrity of their data. The availability and accuracy of agency specific identifiers enable their use as keys in match merges for that agency's data. Agency specific identifiers are involved in a positive feedback loop; patients need valid IDs to receive services, providers must supply valid IDs to be paid for services, and claims processors require valid IDs to pay claims. All parties have a vested interest in ensuring that IDs are valid and recorded accurately. Therefore, the use of an agency's IDs provides a relatively high confidence of matching when using that agency's data.

Match merging is generally not a good choice for combining data from separate agencies even when a common identifier is available on both sets of data. Omissions and errors are more prevalent with identifiers that are not specific to an agency. Both Medicaid and MH/AOD Agencies collect, but generally do not require SSNs for clients. If they are not required, SSNs are often omitted from, or reported incorrectly on, Medicaid and MH/AOD Agency data. When SSNs are collected, they are not verified by the agencies, so errors are common. As a result, attempts to match merge Medicaid data with MH/AOD Agency data using SSNs are prone to errors. Manual review of these match-merged pairs reveal many records that should be linked but are not, as well as incorrect links made because of erroneous SSNs.

Deterministic Linking

Deterministic record linking combines data through the use of identifying information from more than one file. By using multiple criteria, deterministic record linking can overcome the limitations of match merging. Deterministic linking involves comparing identifying information from each of two files and assigning points for each agreement. Only records with a point total over a predefined threshold are linked. A hypothetical deterministic link might score two records using the following criteria:

- 20 points for a complete SSN agreement, or
  10 points for agreement on the last four digits of an SSNs
- 15 points for an agreement on last name
- 8 points for an agreement on first name
- 5 points for a date of birth agreement
- 1 point for a gender agreement, or
  –10 points if gender does not agree

Higher points reflect higher importance of the criterion: complete agreement on an SSN is more important than partial agreement. This hypothetical procedure might link records with scores of 25 points or more.

Deterministic linking clearly provides an improvement over match merging. Errors or absence of SSN, for example, do not prohibit deterministic linking of Medicaid data with MH/AOD Agency data as would be the case with match merging. Record linkage is possible even when an SSN is not available, and errors in an SSN do not necessarily cause incorrect links because SSN agreement alone is not sufficient evidence for linking.

Problems with deterministic linking arise from the difficulty of establishing appropriate points for individual agreement criterion and in setting an appropriate threshold for linking. Points and thresholds cannot be set empirically. In deterministic linking, the number of points awarded for an agreement is arbitrarily set, often through trial and error. While it may be obvious that complete agreement on an SSN should be more important than agreement on the date of birth, it is not intuitive that it is exactly four times as important. Clearly, the relative weighting of the criteria as reflected in the assignment of points is key to the success of the process. This topic is discussed in detail later in the paper.

Points should reflect the relative importance of an agreement. However, the relative importance of an identifier will vary from case to case. Consider the use of Date of Birth as a key in linking two files. Agreements are more important when the universe of potential key values is large than when it is small. When linking two files of high school sophomores, date of birth will not be as important as it is when linking two files of county residents. The reason is that with fewer discrete dates in the key space, chance agreements are more likely.

Another limitation is that deterministic linking does not provide a mechanism for scaling agreement points. For many identifiers, the relative importance of an agreement depends on the value. Consider comparisons of last names. Agreement on a relatively rare name such as "Wobbe" should receive more points than agreement on a relatively common name such as "Smith".

Probabilistic Linking

Probabilistic linking is very similar to deterministic linking – it combines data using identifying information from both files. Like deterministic linking, it uses multiple criteria and scores to establish record links. The difference lies in the manner in which points and thresholds are set. With deterministic linking, agreement points and linkage thresholds are set outside of and known prior to the linking process. This is not the case with probabilistic linking. Agreement points, referred to as "weights", are determined by the data; these are scaled, relative to the value of the identifier. Basing weights on the data creates a flexible method that adapts to differing conditions and overcomes the main weakness of deterministic linking – the arbitrary and rigid assignment of agreement weights.

Probabilistic linking also makes use of disagreements in linking records. This is a convention of probabilistic linking more than a unique characteristic since there is nothing that would preclude the use of disagreements in deterministic linking. The deterministic scoring example above included disagreements for gender. For the most part, however, disagreements are ignored with deterministic linking, while probabilistic linking uses both agreements and disagreements for all identifiers.

While probabilistic record linking resolves some problems of both match merging and deterministic linking, it does so at the cost of complexity. Probabilistic linking involves multiple, non-trivial steps to calculate weights, set linking thresholds, and link the data. The sections that follow explain the details of probabilistic linking in general and the specifics of applying the method to the Medicaid and MH/AOD Agency data for the IDB.

Details of Probabilistic Linking

To provide a framework for understanding probabilistic linking, it is useful to consider the task of match merging files *A* and *B* shown in Table 3 below. Each file contains a key (ID) along with other information (Var1 and Var2). Combining the two files by match merging on the key variable ID yields the results shown in Table 4. The IDs 1 and 3 appear on both File *A* and File *B* and are linked: ID 1 in File *A* is linked with ID 1 in File *B*, ID 3 in File *A* is linked with ID 3 in File *B*. ID 2 appears only on File *A* but not on File *B*, while ID 4 from File *B* does not appear on File *A*. No links were made for either ID 2 or ID 4 because neither was found on *both* File *A* and File *B*.

Table 3: Two Files

File *A*		File *B*	
ID	Var1	ID	Var2
1	E	1	X
2	F	3	Y
3	G	4	Z

Table 4: Files *A* and *B* Combined

ID Links Only		
ID	Var1	Var2
1	E	X
3	G	Y

The next example accomplishes the same match merge, but it uses a different series of steps. This process is not the most efficient way to perform a match merge, but it is illustrative and creates a framework for discussing the process of linking. The steps are:

- Join the two files, creating the Cartesian product of all possible record-pairs,
- Evaluate the IDs of the record-pairs using the match merge rule, and
- Reduce the record-pairs to those pairs where the match merge rule is true.

The first step is creating all possible record-pairs. This involves joining every record from File *A* above with every record from File *B* above as shown in Table 5.

Table 5: File *A* & *B* Record-pairs

Combined Files (Cartesian Product)			
A.ID	*B*.ID	Var1	Var2
1	1	E	X
1	3	E	Y
1	4	E	Z
2	1	F	X
2	3	F	Y
2	4	F	Z
3	1	G	X
3	3	G	Y
3	4	G	Z

Applying the match merge rule for the second step involves comparing the ID from File *A* with the ID from File *B*. If the value of ID in File *A* (*A*.ID) is equal to the value of ID in File *B* (*B*.ID), then the result of the comparison is 1 (or true). If the ID values are not equal, then the result is 0 (or false). This rule can be expressed as a function of the two ID variables:

$$f\,(A.\text{ID}, B.\text{ID}) = 1 \text{ if } (A.\text{ID} = B.\text{ID}), 0 \text{ otherwise}$$

The result of applying this rule to each of the record-pairs is shown in Table 6 below.

Table 6: Record-pairs with Match Merge Rule

Combined Files (Cartesian Product)				
f (*A*.ID, *B*.ID)	*A*.ID	*B*.ID	Var1	Var2
1	1	1	E	X
0	1	3	E	Y
0	1	4	E	Z
0	2	1	F	X
0	2	3	F	Y
0	2	4	F	Z
0	3	1	G	X
1	3	3	G	Y
0	3	4	G	Z

Table 6 adds the result of applying the match merge rule to Table 5. The first column contains the result of the match merge rule: 1 where *A*.ID is equal to *B*.ID and 0 where the values are not equal. By keeping only rows where the match merge rule, *f* (*A*.ID, *B*.ID) returns one – or true –, we obtain the equivalent of Table 4.

We can extend this framework to describe other procedures for deterministic and probabilistic linking. Where the preceding match merge example uses a single comparison key (ID), linking uses a collection of identifying variables such as name, SSN and date of birth. Where match merging uses a test of equality to determine whether records should be combined, linking employs a more sophisticated comparison function incorporating several weight-based comparisons. The score derived from these comparisons is evaluated against a threshold to determine links.

Blocking

Blocking is the process of creating record-pairs only when there is some evidence for linking the two records. The idea is to eliminate record-pairs that lack evidence for a link, resulting in a smaller decision space and a more efficient search for links (Newcombe, 1967). The initial decision space for linking two files is the Cartesian product of those two files. For files of roughly equal size, the size of the decision space increases exponentially with the size of the files while the proportion of record-pairs that potentially represent links decreases as shown in Table 7 below. Blocking decreases the size of the decision space, thus reducing the overall number of comparisons to be made.

Table 7: Potential Links as a Percentage of the Decision Space

			Potential Links	
Records in File *A*	**Records in File *B***	**Decision space**	**Number**	**Percent of Decision Space**
100	100	10,000	100	1.000%
1,000	1,000	1,000,000	1,000	0.100%
10,000	10,000	100,000,000	10,000	0.010%
100,000	100,000	10,000,000,000	100,000	0.001%

Identifying Variables

The linking process requires the use of identifying variables – variables that can be used to identify a person. Some variables are strong identifiers. An SSN is an example of a very strong identifier: in theory, everyone should possess a single SSN and every SSN should apply to one and only one person. Unfortunately, this is not always the case. A name is also a strong identifier. While several individuals may share a particular name, the name is still a useful, although not perfect, means of identification. Other variables are weak identifiers, but they are still useful in identifying a person. A person's gender is an example of a weak identifier: it provides some, but not much, discriminating power. A person's gender can help identify a person when used in conjunction with other identifiers, such as date of birth and name.

In order to use identifying variables as comparison variables, they must be usable and overlapping. Identifying variables that contain values are usable. A variable that is blank for all or most records on a file is not usable. An identifying variable that is usable *on both files* is an overlapping variable – that is, it is available on both files. As noted above, an SSN is a very strong identifier, but it cannot help with linking if it is available only on one of the two files.

Comparisons

Comparisons are part of the process of evaluating record-pairs to determine links. Each pair of comparison variables is evaluated. The extent of agreement or disagreement between the two variables is the result of the comparison. Comparisons can be dichotomous or continuous as explained below.

Dichotomous comparisons are unforgiving of errors and mistakes, there are no gray areas. Comparisons of this type are generally structured to use values of zero or one (0 or 1) to indicate disagreement or agreement respectively. A gender comparison is an example of a dichotomous comparison – the genders are either the same or they are different.

Continuous comparisons allow for partial agreement, returning a result that lies on a numeric scale ranging from disagreement to agreement. In the case of names, a continuous comparison applied to a misspelling can indicate that the names are partially in agreement. Continuous comparisons indicate not just agreement, but the degree of that agreement.

While every comparison can be expressed with a dichotomous result, continuous comparisons often provide more useful information. Consider comparisons of names. Misspellings and nicknames cause problems with dichotomous comparisons. With misspellings and nicknames, dichotomous comparisons evaluate as disagreements even when the two records represent the same person. One technique used to overcome this problem is the use of phonetic equivalents such as the Russell Soundex code or the New York State Identification Information System (NYSIIS). Phonetic equivalent techniques work by grouping letters with similar sounds and suppressing vowels. They can overcome some misspellings, but at the cost of hiding differences that may be important. A Soundex comparison that evaluates as an agreement does not necessarily indicate that the underlying names are identical.

For the IDB, phonetic comparisons were not made. A process known as approximate string matching was used instead – this is described in a later chapter. NYSIIS phonetics were used to group names for calculating and using scaling factors. The `#nysiis.sas` module includes code for the IDB's implementation of this algorithm.

Weights

The discriminating power of each comparison variable – its importance in determining links – is expressed as a weight. Weights reflect the probability that agreements and disagreements occur by chance. Weights are calculated from the relative probabilities of agreement (and disagreement) for the linked and non-linked record-pairs, and can be expressed as

$$Weight = \log_2\left(\Pr_{link} \big/ \Pr_{non-link} \right)$$

where $\Pr_{link}$ is the probability in the linked record-pairs, and $\Pr_{non-link}$ is the probability in the non-linked record-pairs.

Agreement and Disagreement Weights

For each comparison variable, there are two weights: an agreement weight and a disagreement weight. Both agreements and disagreements are important in determining links. Agreement and disagreement weights are not symmetrical – a disagreement weight is not equal to the negative of the agreement weight.

Weights for gender demonstrate this quality. Gender is not a strong identifier and does not provide much evidence by itself that the two halves of a record-pair should be linked. Disagreement on gender, however, is a strong indicator that the two halves of a record-pair should not be linked. Because gender is either male or female (ignoring the few records without gender information), there are many gender agreements in the decision space. Among non-links, there is gender agreement with approximately half the record-pairs. This contrasts with the links, where gender agreement occurs with nearly all the record-pairs. Actually, gender disagreement is found among links only where gender-recording errors are encountered. As the accuracy of the data increases, the agreement weight approaches the number positive one, while the absolute value of the disagreement weight grows larger. This can be easily demonstrated using the weight formula shown above.

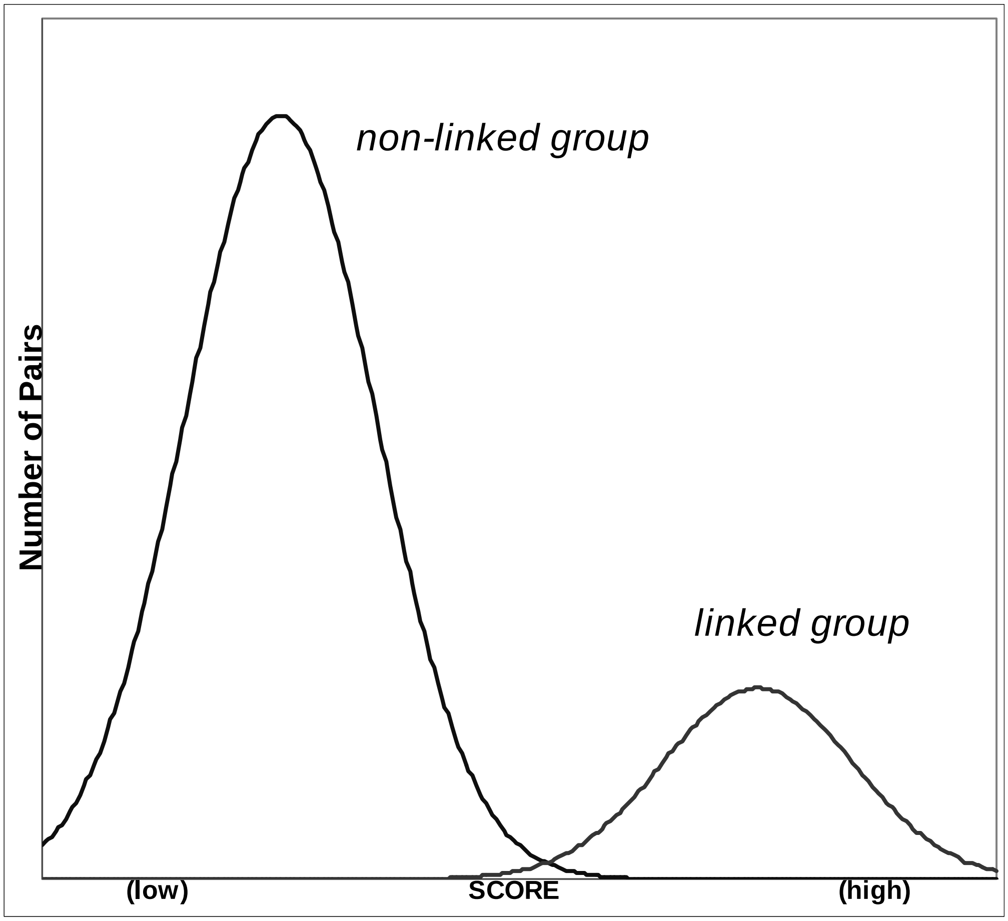
Scaling Factors

Weights reflect a variable's overall importance for agreement (or disagreement) over the entire decision space. The weight for last name agreement is calculated based on the relative probabilities of agreement for links and non-links. As a result, there is one last name weight. As stated above, weights should reflect not only the probability that an agreement (or disagreement) happens by chance, but also the importance of an agreement. For example, agreement on a common name such as "Smith" should be less than agreement on an uncommon name, such as “Wobbe”. Scaling provides a mechanism for modifying a weight based on the relative frequency of a comparison variable's value. If the last name "Smith" was frequently found in the decision space, then the scaling factor for "Smith" would be used to adjust the last name weight down, or lower on comparisons involving that name. Conversely, if "Wobbe" were found infrequently among last names in the decision space, the scaling factor would adjust the last name weight up, or higher.

Scoring

Scores for record-pairs that should be linked will vary, as will the scores for the record-pairs that should not be linked. There will be high and low scores for each group, and these scores will be distributed somewhat normally within each group. Generally, scores for the linked group will be higher than scores for the non-linked group. Figure 1 shows the hypothetical distribution of a non-linked group and a linked group.

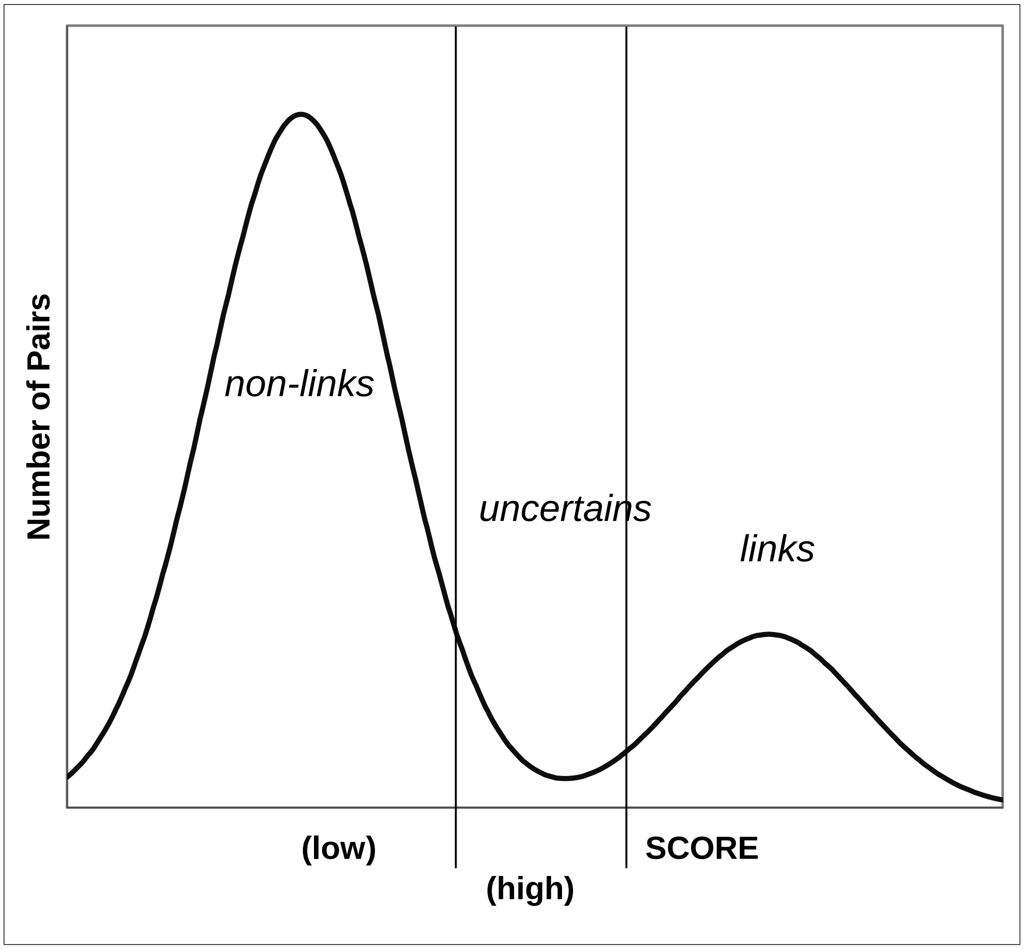
Figure 1: Scores for Linked and Non-linked Record-pairs



Prior to the actual linking, it is not known which record-pairs should be linked and which should not. There is only one group of record-pairs. This one group is a combination of link and non-link record-pairs. Some of the record-pairs have high scores and some have low scores – scores determined from comparisons of individual variables and their associated weights. Overall, these scores will have a bimodal distribution. With the knowledge that scores for record-pairs that should be linked are generally higher than scores for the record-pairs that should not be linked, the record-pairs can be classified as links, non-links, and uncertain pairs as shown in Figure 2.

Probabilistic record linking results in a range of scores for record-pairs. Record-pairs at the ends of the distribution represent links and non-links, but scores in the middle fall into the uncertain category and the uncertain record-pairs must still be reviewed manually. If the process is correctly set up and good data are available, the number of uncertain record-pairs should be of a manageable size for human (non-automated) review.

Figure 2: Linked, Non-linked, and Uncertain Record-pairs



Determining Weights

Probabilistic linking weights are unique for each combination of joined files; recalculation is necessary for each new set of files. The explanation of weight calculations above contains some circular logic: the weight calculations need the record-pairs divided into links and non-links. But the record-pairs have not yet been categorized into links and non-links prior to determining the weights.

To solve this problem, weights are typically calculated manually from file samples. A sample of records from one file is joined with a sample of records from a second file, comparisons are made, and then the joined records are manually reviewed to determine links. Only after links are determined, is it possible to separate the sample decision space by links so that weights can be accurately calculated. The size of the sample used to create the weights is the key to the accuracy of the weights that are developed. Larger samples require more time to evaluate, but produce weights that are more accurate and generally create fewer uncertain record-pairs. This reduces the amount of time necessary for manual review. Conversely, smaller samples tend to reduce the time needed to determine weights, but usually increase the number of uncertain links and the time necessary for manual reviews. Therefore, while probabilistic linking produces results superior to those of other methods, the time factor can make the process impractical. For the IDB, however, we have developed an iterative linking process that develops accurate weights and requires neither sample files nor an initial manual linking step. The process starts by joining the data, performing an initial deterministic link, and then calculating probabilistic weights using the deterministic classifications. The following chapter describes this process and other specific techniques used on the IDB project.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter Four: The IDB Probabilistic Linking Process

Creating the IDB by integrating data from multiple sources involves two related issues: linking the data sources, and unduplicating individuals within a single source of data. The obvious issue is linking two data sources, but accurate linking assumes unduplicated data sources. Consider the situation where file *A* contains one record for John Smith and file *B* contains two records for the same John Smith. Linking associates the two John Smiths in file *B* with the John Smith in file *A*, in effect it unduplicates file *B*. In this example, linking file *A* with file *B* assumed that file *A* did not contain any duplicated records. Had file *A* contained two records for John Smith, the results of linking the two files would have been ambiguous. To successfully link two data sources, at least one of the data sources must not contain any duplication. One approach to this problem is to select a data source that is assumed to be free of duplication. This is often a risky assumption. Another approach is to unduplicate one of the data sources prior to linking with other data sources. The IDB made use of a third approach that combines unduplication with linking.

Combining unduplicating with linking reduced the number of steps required to link the MH/AOD Agency and Medicaid data. This in turn reduced the number of reviews and the amount of time necessary to complete the process. Had the combined approach not been used, the process would take the following progression:

- Unduplicate the Medicaid client records,
- Link the (combined) MH/AOD Agency client records to the unduplicated Medicaid client records,
- Unduplicate any MH/AOD Agency client records that did not link to any Medicaid client records.

This would be a time intensive process because each of these steps requires a manual (non-automated) review to resolve the uncertain record-pairs.

Combining the unduplicating and linking steps reduced the amount of manual review required. This was possible because unduplicating a single data source is very similar to linking two (or more) data sources. Unduplicating a single file, in fact, is actually the process of linking a file with itself. Both processes (unduplicating and linking) use the same procedures and involve the same evaluation criteria. Based on this knowledge, the unduplicating and linking process described above was combined for the IDB:

- Combine (concatenate) the Medicaid client records with the MH/AOD Agency client records,
- Unduplicate the combined client data.

This unduplicating process is a specialized linking process, and for the remainder of the report, we will refer to it as linking.

The entire IDB linking process consists of three steps:

- IDB Probabilistic linking,
- Manual Review of Uncertain Record Pairs, and
- Mapping of Identifiers,

Descriptions of these steps follow, in the order they were performed. All linking programs and subroutines are listed in the Appendix of this report and are available from the project web site (http://www.samhsa.gov/centers/csat/content/idbse). Requests for the specific SAS programming code for each step can also be directed to SAMHSA's Center for Substance Abuse Treatment, Office of Managed Care at (301) 443-8796. The IDB process consists of two programs separated by a manual review step. This process was successfully applied to all three States.

Step One - IDB Probabilistic Linking

The first step, contained in program `step350.sas`, creates the decision space, calculates weights, and separates the record pairs into links, non-links, and uncertains. This step consists of several enumerated sections and is described below.

1. Concatenating the Identifying Data
2. Calculating the Scaling Factors
3. Blocking, Joining, and Comparing the Data
4. Deterministic Linking and Initial Weight Calculations
5. Probabilistic Linking Iterations
6. Final Probabilistic Linking Iteration
7. Printing the Uncertains for Manual Review

1. Concatenating the Identifying Data

The first linking step combines, or concatenates, the MH/AOD Agency and Medicaid client data into a single data set. The MH/AOD Agency data may reside in one or more files. The combined data will be joined to itself to determine links. Code for concatenating the client data is included in the `step350.sas` program.

2. Calculating the Scaling Factors

The combined data are also used to calculate scaling factors (described in the previous chapter). Scaling factors are associated with specific values of variables and are inversely related to the relative frequency of that value. Scaling factors for common values reduce weights, while scaling factors for uncommon values increase weight values. With scaling, it is possible to calculate overall agreement and disagreement weights for last name while recognizing that agreements on uncommon values are more important than agreements on common values.

Scaling is performed for comparison variables where an even distribution of values is not expected. An SSN variable, for example, is not scaled because we expect an even distribution of SSN values. Comparison variables that are scaled include:

- First Name (based on NYSIIS phonetic)
- Middle Initial
- Last Name (based on NYSIIS phonetic)
- Birth Year
- Race
- ZIP Code

Scaling factors are inversely related to the relative frequency for a particular value. Code for assigning scale values is included in `#scale.sas`.

3. Blocking, Joining, and Comparing the Data

After the scaling factors are calculated, the data are blocked and joined to create the decision space for the linking process. With the decision space available, all comparisons are made. Results of these comparisons are retained throughout the remainder of the process. IDB linking is an iterative process, but rather than repeating the comparisons with each iteration the comparisons are made once at the start of the process. The high level code for this section is saved as `#joindata.sas`.

Data Blocking and Join

`#joindata.sas` creates the decision space by joining the combined data to itself. Blocking is used to eliminate many of the non-links from the decision space, not to restrict links. Multiple, overlapping blocking criteria are used at this stage – a record-pair need satisfy only one of these criteria for inclusion in the decision space:

- Agreement on SSN,
- Agreement on date of birth and phonetic (NYSIIS) last name,
- Agreement on date of birth, gender, and phonetic (NYSIIS) first name, or
- Agreement on gender, phonetic (NYSIIS) first name, and phonetic (NYSIIS) last name.

The record joins are performed using code in `#join.sas`.

Comparisons on Identifying Variables

As noted above, comparisons are made once at the start of the process for all comparison variables. This is done because some of the comparisons require a great deal of computing resources. Results of the comparisons are retained throughout the process. The MH/AOD Agency and Medicaid data used for the IDB contained an abundant set of identifying variables and there was sufficient overlap between the two sources. Ten comparisons – using IDs, names, and demographic information – were made in determining IDB links. These comparisons are discussed below. Code for performing the comparisons can be found in `#compare.sas`.

IDs and Names

A rich set of names and IDs are available for comparisons from data sources that contribute to the IDB. In addition to first and last name, the Medicaid data also contain a field for maiden name, although this is not always available. Where a maiden name is available, it is used as an alternative for last name comparisons. If there is no agreement on last name, then the MH/AOD Agency last name is compared with the Medicaid maiden name. The list of names and IDs used in linking IDB data include:

- SSN
- Medicaid ID
- First Name
- Middle Initial
- Last Name
- Maiden Name

An SSN variable was included on all data sources for the IDB, and reasonable values were available on 60 to 80 percent of the records depending on the data source. The Medicaid ID was available for all Medicaid clients. Most MH/AOD Agency data also included a Medicaid ID field, although it was frequently not used. Exceptions were Washington AOD Agency data and Delaware adult MH/AOD Agency data. Medicaid IDs found on the remainder of the MH/AOD Agency data often contained questionable values.

Name comparisons code is found in `#ncomp.sas`, while the code for ID comparisons is found in `#compare.sas`. Both names and IDs were compared using a technique known as *approximate string matching*. Approximate string matching is a continuous comparison that calculates the percentage of agreement between two strings. The methodology subtracts the number of additions, deletions, and changes necessary to force complete agreement divided by the length of the longer string from one (Landau & Vishkin, 1989). To prevent misleading results, comparisons of less than 70 percent agreement were reclassified as disagreements.

For example, approximate string matching the names "Johnson" and "Johnston" requires either the addition of a "t" to "Johnson", or the deletion of "t" from "Johnston". The methodology subtracts 1 divided by 8 (the length of "Johnston") from 1, with the result of 0.875. In other words, there is an 87.5 percent agreement between "Johnson" and "Johnston". Approximate string matching code is in `#asm.sas`.

Demographic Information

The following demographic information, used in linking the AOD and Medicaid data, was found on most IDB data sources.

- Date of Birth
- Race
- Gender
- ZIP Code

Exceptions include Delaware adult MH/AOD Agency and Washington Community and Institutional MH Agency data. The ZIP code is not available on either the adult Delaware Psychiatric Hospital data nor the Washington Community or Institutional MH Agency data.

Dichotomous comparisons were used for both race and gender, evaluating only exact matches as agreements. Date of birth and ZIP codes were compared using continuous comparisons. If a date of birth comparison was not a complete agreement, but two of the three date components (year, month, and day) were in agreement, then a partial agreement was calculated. The partial agreement was calculated as the difference in days between the dates divided by the age in days for the Medicaid date. The file `#dobcomp.sas` contains the code for comparing birth dates. For five digit ZIP code comparisons, partial agreements were calculated if there was not an exact match. Partial agreement for ZIP codes was calculated as the distance in miles between the centroids of the two ZIP codes, divided by a State-specific constant and subtracted from 1.0. The ZIP code comparison is contained in the file `#zipdist.sas`.

4. Deterministic Linking and Initial Weight Calculations

The next phase makes an initial deterministic link that is used in calculating the first set of probabilistic weights. Code for this phase is contained in the file `#prelimwt.sas`. The deterministic link is performed on the data to make an initial determination of links and non-links. Multiple deterministic criteria are used:

- Agreement on SSN, Medicaid ID, date of birth, and gender.
- Agreement on SSN, date of birth, and gender – with partial agreement on one name variable (80 percent on first name or 90 percent on last name or complete agreement on middle initial).
- Agreement on Medicaid ID, date of birth, and gender – with partial agreement on one name variable (80 percent on first name or 90 percent on last name or complete agreement on middle initial).

- Partial agreement on name (80 percent on first name and 90 percent on last name) plus agreement on date of birth, gender, and either ZIP Code or race.
- Partial agreement on name (80 percent on first name and 90 percent on last name) plus complete agreement on date of birth and partial agreement (90 percent) on either SSN or Medicaid ID.
- Partial agreement on name (80 percent on first name and 90 percent on last name) plus complete agreement on date of birth and middle initial.

These criteria determine the initial link/non-link classifications, which are used to prepare the first probabilistic weights and thresholds. This involves the following steps:

- Calculate weights using the `#weight.sas` code,
- Compute scores for record-pairs using the new weights and the scaling factors –code for these is found in `#prelimwt.sas` and `#scalewt.sas`,
- Determine thresholds for dividing the decision space into links, non-links, and uncertains - using the record-pair scores and the deterministic link classifications (this code is found in the file `#thrshld.sas`).

5. Probabilistic Linking Iterations

With the initial probabilistic weights in hand, the next section of the `step350.sas` program performs the first probabilistic linking iterations. Code for this phase is found in the file `#iterwts.sas` and includes the following steps:

- Reclassify the decision space using the record-pair scores and thresholds from the previous phase – uncertains are reclassified as links or non-links with the deterministic criteria described above (this code is found in the file `#mtchcls.sas`),
- Recalculate weights using the `#weight.sas` code,
- Recompute scores for record-pairs using the new weights and the scaling factors –code for these is found in `#iterwts.sas` and `#scalewt.sas`,
- Determine thresholds for dividing the decision space into links, non-links, and uncertains - using the record-pair scores and the deterministic link classifications (this code is found in the file `#thrshld.sas`).

6. Final Probabilistic Linking Iteration

Code for the final probabilistic link is found in the file `#finalwts.sas`. This final link divides the record-pairs into links, non-links, and uncertains using the record-pair scores and thresholds from the previous phase.

7. Printing the Uncertains for Manual Review

The final section of the program `step350.sas` prints a list of the uncertain record-pairs. This report is used in the next step of the IDB probabilistic linking process, the manual review. The report code is contained in `#review.sas`.

Step Two - Manual Review of Uncertain Record-Pairs

The report from `step350.sas` lists all uncertain record-pairs for manual review. The number of such record-pairs varied by state – the number of uncertains for each state is discussed in a subsequent chapter. The manual review selects all uncertain record-pairs that represent links. The selected record-pairs augment the links determined by `step350.sas`, which are used by the second program to map IDs.

Step Three - Mapping of Identifiers

The third and final linking step involves the program, `step352.sas`, which processes the linking results from prior two steps and assigns a new IDB identifier to all client records. It then maps the MH/AOD Agency IDs and Medicaid IDs to the new identifier. The ID mapping from this step is used in constructing the final IDB files. The phases for this step are:

- Assign IDB identifiers for linked record-pairs,
- Assign IDB identifiers for clients not linked,
- Create a master file of identifying variables,
- Create files for mapping original IDs to the new IDB identifiers.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter Five: Comparison of Linking Results

Programs developed for the IDB successfully linked client records for three States using probabilistic record linking techniques. Processing and review time varied depending on the volume of data, the availability of comparison variables, and the quality of the comparison variables. Table 8 shows the approximate resources used in linking the data for each of the three States. The Delaware files were the smallest of the three States and required the fewest resources to link. Washington files, in contrast, were the largest and required longer to process and review than any of the other States. In addition to the larger files, an important contributor to the high resource usage in linking the Washington data was probably the absence of some comparison variables on the MH Agency data.

Table 8: Linking Resources Used

State	MH/AOD Agency Records	Medicaid Records	Approximate Processing Time	Uncertain Record-Pairs	Approximate Review Time
Delaware	13,500	101,713	10 minutes	794	4 hours
Oklahoma	119,000	467,176	90 minutes	1,892	9 hours
Washington	138,447	1,335,387	400 minutes	4,017	19 hours

Results of the linking were far superior to those for a simple match merge based on SSN. Table 9 shows the number of IDs linked between MH/AOD Agency data and Medicaid IDs for each of the three States. For all States, the number of links was greater than the number from the match merge. Overall, the IDB process created over 80 percent more links than would otherwise be possible. The increase ranged from 20 percent for the Delaware data to nearly 150 percent for Washington MH Agency data. More important, the probabilistic process discovered links not found with the match merge and eliminated false links generated by the match merge.

Table 9: Comparison of Links from Probabilistic Linking and Match Merge on SSN

(1) State	(2) Probabilistic Linking Total	(3) Links Not Found with Match Merge	(4) Links Common to Both Probabilistic and Match Merge	(5) False Links from Match Merge	(6) Match Merge Total
Delaware	5,612	1,488	4,124	523	4,647
Oklahoma	38,683	18,655	20,028	6,809	26,837
Washington AOD-Medicaid	14,836	4,766	10,070	588	10,658
Washington MH-Medicaid	56,990	35,135	21,855	995	22,850
Washington AOD-MH	7,830	6,562	1,268	2,412	3,680

The third and fifth columns of Table 9 represent "errors" created by using a match-merge technique to link data. There are two types of errors associated with any linkage methodology: false positive and false negative links. *False positive* links are record-pairs that should not be linked, but which were deemed links by the methodology. *False negative* links are record-pairs that should be linked, but were not linked by the methodology. For links determined by the match-merge routine, column three of Table 9 ("Links Not Found with Match Merge") represents false negative links whereas the fifth column ("False Links from Match Merge") represents false positive links. These are errors associated with match-merged links.

Virtually all linking schemes include some level of error. More sophisticated techniques tend to reduce the overall error level, but they do not eliminate it. Given some level of error, there is a tradeoff between false positive and false negative links; lessening one tends to increase the other. The linking philosophy for the IDB was to minimize the number of false links in the database with the knowledge that the database would then contain false negatives – separate clients that actually represent one person.

The tables that follow show the number of IDs from the MH/AOD Agency data and the number of those IDs that were linked with a Medicaid ID for each State. The IDB linking process generated ID overlaps of 30 to 60 percent depending on the files.

Delaware

More than 40 percent of the Delaware MH/AOD Agency clients were linked with a record in the Medicaid data. Table 10 shows the overlap between the MH/AOD Agency clients and the Medicaid clients in the Delaware IDB. Any overlap between the adult and children's data was not examined. The populations served by the two MH/AOD Agencies should overlap only for those persons reaching majority during the study year, so the overlap between the two agency populations should be minimal.

Table 10: Delaware MH/AOD Agency-Medicaid Client Overlap

Unique Clients from the MH/AOD Agency Data	13,460
MH/AOD Agency Clients Linked with Medicaid Clients	5,612
Percentage of MH/AOD Agency Clients Linked	41.7%

Oklahoma

Nearly one third of the Oklahoma MH/AOD Agency clients were linked with records in the Medicaid data as shown in Table 11.

Table 11: Oklahoma MH/AOD Agency-Medicaid Client Overlap

Unique Clients from MH/AOD Agency Data	118,962
MH/AOD Agency Clients Linked with Medicaid Records	38,683
Percentage of MH/AOD Agency Clients Linked	32.5%

Washington

Two different Washington agencies supplied MH/AOD Agency data for this project. Since it was anticipated that the AOD and MH Agencies would serve overlapping populations, the overlaps were examined separately.

The overlap between AOD Agency and Medicaid clients is shown in Table 12. Over 38 percent of the AOD Agency client IDs were linked with records in the Medicaid data.

Table 12: Washington AOD Agency-Medicaid Client Overlap

Unique Clients from AOD Agency Data	38,538
AOD Agency Clients Linked with Medicaid Records	14,836
Percentage of AOD Agency Clients Linked	38.5%

Nearly 58 percent of the MH client IDs were linked with records in the Medicaid data as shown in Table 13.

Table 13: Washington MH Agency-Medicaid Client Overlap

Unique Clients from MH Agency Data	98,533
MH Agency Clients Linked with Medicaid	56,990
Percentage of MH Agency Clients Linked	57.8%

Table 14 shows the overlap between the AOD and MH Agency clients. Over 20 percent of the AOD Agency client IDs linked with a MH Agency ID (approximately 8 percent when expressed as a percentage of the MH Agency IDs).

Table 14: Washington AOD-MH Client Overlap

Unique Clients from AOD Agency Data	38,538
AOD Agency Clients Linked with MH Agency Clients	7,830
Percentage of AOD Agency Clients Linked	20.3%
Unique Clients from MH Agency Data	98,533
MH Agency Clients Linked with AOD Clients	7,830
Percentage of MH Agency Clients Linked	7.9%

Chapter Six: Conclusion

The IDB links physical health services and MH/AOD services provided by Medicaid with specialized MH/AOD services provided by MH and AOD Agencies and should prove a valuable resource for the research and evaluation of the MH/AOD population. The integration effort reveals service usage across each of the three systems and provides a more complete picture of the health needs of MH/AOD Agency clients. The comprehensive view of service utilization provided by the IDB can:

- Identify comorbidities, both across MH and AOD and across medical conditions – for example, diabetes and depression,
- Analyze utilization by region and demographics – highlighting under-served populations,
- View utilization levels across agencies – facilitating the coordination of care at the macro level.

One of the requirements of the IDB project was to create procedures for accurately linking person-level records from multiple files with a minimum of human intervention. This report examines the process developed by the project team and the results achieved with that process. It is our conclusion that the probabilistic linking process employed on the Integrated Database Project successfully and efficiently linked large amounts of disparate data from a variety of sources while minimizing error introduced into the process and the requirement for manual review and correction. Compared with more conventional match merge and deterministic linking methods, the IDB procedures created more links and more accurate links.

We believe the process can be easily adapted to perform linking between Medicaid data and MH/AOD Agency data for other States, and with modest effort, it can be applied to other similar health service research linking tasks. The probabilistic linking process is relatively complex (compared to simple match merging) and involves a number of steps, but the benefits of using the process are significant. Alternatives are to accept the errors introduced by a simpler process, or to employ more human effort in the linking task. Neither of these alternatives is very appealing. The error rate for match merging can be moderate to high, depending on the availability and quality of the information employed. Manual linking, which can achieve very good accuracy, can be employed for small volumes of data measured in the hundreds or thousands of individuals. The IDB probabilistic linking methodology relies on manual linking only as the last step in the process to resolve a small number of ambiguous cases. However, manual linking quickly becomes impractical as a primary linking method as the volume of data increases. Clearly, for the task we faced in building the IDB, and the task faced by States that want to replicate this methodology, manual linking is not an option.

There is some level of error associated with every methodology. The overall level of error can be reduced, however, through improvements and refinements in methodology. Probabilistic linking provides better results than either match merging or deterministic linking. With any linking methodology, there is a tradeoff between false positive and false negative links; methodological modifications that may lessen one result tend to increase the other result. The approach taken for the IDB was to favor false negative error over false positive error. This choice was made deliberately to minimize the number of false links in the database with the assumption that false negative links would have a more detrimental effect on typical research objectives than false positives.

The results of linking were fairly consistent across States, and they provide a very interesting basis for further exploration. Values ranging from 38 percent to 58 percent of the MH and AOD Agency clients were linked to each other and to Medicaid client data. In the Washington data a surprising 58 percent of the MH client IDs were linked with corresponding Medicaid IDs (Table 13).

Record linking requires a considerable amount of computing power and data storage. Nevertheless, the results were achieved with an expenditure of machine resources that is very reasonable considering the difficulty of the task. These results were not achieved by brute force supercomputing methods. Four years ago, during the planning stages of the IDB project, the computer resources necessary to link the IDB data and build the databases were beyond the reach of PC based systems. The equipment ultimately selected includes a 64 bit, 400 MHz processor and more than 100 gigabytes of storage – impressive at the time. Today this type of power is available with many high-end PC based servers. The power of the machine used to perform the linking is very much within the reach of States, and we see few technical impediments to the dissemination and utilization of the IDB probabilistic linking methodology.

The IDB process links person-level data and was designed in a modular manner to facilitate modifications. All processing is performed using The SAS System and all the programs and modules are SAS programs. These programs can be run on a wide array of platforms, from Windows-based PCs through server-class Unix machines, to high-end mainframes. The programs and modules, listed in the appendix, are available on the CSAT web site at http://www.samhsa.gov/centers/csat/content/idbse/idbsas.asp. The programs are available as a resource to other organizations interested in linking person-level data. Modifying the programs to work with other data should require two to three weeks of programming – depending on the extent of the changes. The time required to perform linking and review time will vary according to the number of unique client records, although the information in Table 8 can be used to estimate these requirements.

Future work will focus on adapting the linking process to add additional years of data to the IDB databases. Considerations include managing ID and name changes over time, and possibly, incorporating new data sources. With a review time of approximately 19 hours for the State of Washington, the manual review process will have to be very carefully considered and perhaps refined for larger States, such as California, that want to perform linking. However, the linking process is still well within the realm of possibility for larger states.

References

Fellegi, I.P., & Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183-1210.

Landau, G.M., & Vishkin, U. (1989). Fast Parallel and Serial Approximate String Matching, *Journal of Algorithms*, *10*, 157-169.

Newcombe, H.B. (1967). Record linking: The design of efficient systems for linking records into individual and family histories. *American Journal of Human Genetics*, *19*(3) 335-359.

Newcombe, H.B., Kennedy, J.M., Axford, S.J, & James, A.P. (1959). Automatic linkage of vital records. *Science, 130*, 954-959.

Romano, P.S., & Luft, H.S. (1992). Getting the most out of messy data: Problems and approaches for dealing with large administrative data sets. In M.L. Grady (Ed.) *Medical Effectiveness Research Data Methods* (AHCPR Pub. No. 92-0056). Washington, D.C.: U.S. Government Printing Office.

U.S. gazetteer place and zipcode file [Electronic data file]. (2000). Washington, D.C.: United States Census Bureau. Available: http://www.census.gov/ftp/pub/tiger/tms/gazetteer/zips.txt.

Weber, G.I. (1995). Achieving a patient unit record within electronic record systems. In Medical Records Institute (Ed.) *Toward an Electronic Patient Record* (pp. 126-134). Newton, MA: Medical Records Institute.

Appendix: Client Linking Program Description

All linking code for the IDB project was constructed using The SAS System. The pages that follow list the SAS programs and modules developed for the IDB linking programs and modules. The two programs are listed first, followed by the modules in the order first used (some modules are used in several sections). The source code is available from http://www.samhsa.gov/centers/csat/content/idbse. Requests for source code can also be directed to the SAMHSA's Center for Substance Abuse Treatment, Office of Managed Care at (301) 443-8796.

Program/Module Listings

Program `step350.sas -` First step in client linking program. Combines and scores the identification data.

Program `step352.sas -` Second step of client linking program. Assigns manual review pairs with positive matches from step350 and completes the linking of ID files.

Module `#scale.sas` - Creates a data set listing the variables of the supplied data set.

Module `#joindata.sas -` Combines client observations to determine weights and links.

Module `#prelimwt.sas` - Prelimwt.sas determines preliminary weights for linking variables.

Module `#iterwts.sas` - This program develops weights for all variables used in linking.

Module `#finalwts.sas -` Applies the final weights to determine matches

Module `#join.sas -` Combines two data sets using specified criteria and creates a data view of joined data.

Module `#compare.sas -` Combines data and compares identifying variables.

Module `#weights.sas -` Summarizes data by matched/nonmatched and calculates new weights.

Module `#scalewt.sas -` Adjusts weights for first and last names, date of birth, and ZIP codes according to the appropriate scale factors.

Module `#thrshld.sas -` Calculates matched/nonmatched/uncertain threshold from scored data and creates upper and lower bound thresholds.

Module `#mtchcls.sas -` Classifies joined client pairs as matched or not matched using scores and the upper threshold.

Module `#ncomp.sas -` Compares two (arrays) of names and returns a numerical score indicating the degree of agreement.

Module `#dobcomp.sas -` Calculates the relative 'distance' between two birth dates.

Module `#nysiis.sas -` Creates phonetic code for name based upon the New York State Identification and Intelligence System (NYSIIS).

Module #asm.sas - Calculates 'distance' between strings using approximate string matching algorithm.

Module #twinchk.sas - SAS macro to identify high scores resulting from the pairing of twins rather than multiple IDs for the same person.

Module #zipdist.sas - Calculates distance between ZIP codes based upon the longitude and latitude of each ZIP code's centroid.

Module #review.sas - Creates a report for manual review of linked client record pairs.

Module #iddata.sas - Variable attributes for IDMASTER data.

Module #trans.sas - Translates multiple, associated links to a common link.